Prediction of Chronic Kidney Disease Using Machine Learning Techniques

Prabakaran G[1], Harshini S[2], Shalini D[3], Sharumathi K[4], Sowmiya S[5]
[1]Assistant Professor, Dept of Electronics and Communication Engineering
[2, 3, 4, 5]Dept of Electronics and Communication Engineering
[1, 2, 3, 4, 5] Vivekanandha College of Technology for Women, Tiruchengode, India

Abstract- *More persons with advanced stages of renal illness are diagnosed each year.Deteriorating kidney function over time is a hallmark of chronic kidney disease, also known as renal disease. People are more likely to be examined for chronic kidney disease (CKD) if they have diabetes, high blood pressure, or a family history of the illness. An early diagnosis is essential for successful therapy. In order to avoid kidney damage or failure, early detection and monitoring are essential. Healthcare organizations can effectively use machine learning (ML) to support their decision- making process. The goal of this project is to create and suggest a machine learning approach for chronic kidney disease prediction. The Random Forest (RF) algorithm was our suggestion. To determine the optimal model for prediction, the components are constructed using datasets related to chronic renal disease and compared.*

Keywords- Chronic Kidney Disease, Machine Learning.

I. INTRODUCTION

The diagnosis of CKD has improved as a result of their research. In the previously described models, mean imputation is utilized to fill in the missing data based on the diagnostic categories of the samples. Their procedure cannot be employed because the diagnostic results of the samples are unknown. In actuality, patients may fail specific tests for a number of reasons before receiving a diagnosis. Additionally, data obtained using mean imputation may differ greatly from the real values when missing values are included in categorical categories.

1.1 CHRONIC KIDNEY DISEASE

Kidney function steadily deteriorates over a period of months or years in chronic kidney disease (CKD). Disorientation, leg swelling, exhaustion, nausea, and vomiting may ensue after the initial lack of symptoms. Among the complications are heart disease, anemia, excessive blood pressure, and fractures. Chronic kidney disease can be exacerbated by gout, diabetes, high blood pressure, polycystic kidney disease, and other illnesses. One risk factor is a family history of chronic renal illness. The diagnosis is made with a urine test to measure albumin and a blood test to measure the estimated glomerular filtration rate (eGFR). To determine the underlying reason, an ultrasound or kidney biopsy may be performed.

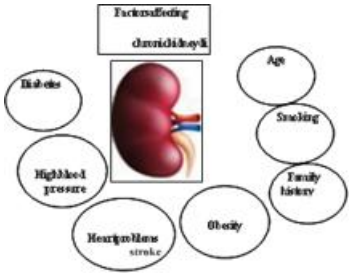

Figure 1. Factors affecting chronic kidney disease

1.2 MACHINE LEARNING

The study of computer algorithms that improve on their own over time is known as machine learning (ML). According to certain theories, it is a subset of artificial intelligence. Machine learning algorithms employ sample data, sometimes known as "training data," to construct a model that allows them to make judgments or predictions without explicit programming. Machine learning algorithms are used in many applications, such as computer vision and email filtering, where it would be difficult or impossible to create classical algorithms that can perform the task. But statistical learning is only one aspect of machine learning. Though there are other forms of machine learning in addition to statistical learning, a subset of machine learning is strongly tied to computational statistics, which is concerned with computer-generated predictions. The study of mathematical optimization provides the theory, methodology, and application areas of machine learning. Data mining, which emphasizes exploratory data analysis through unsupervised learning, is a related topic of research. AI is the process by which computers learn how to complete tasks without explicit training.

II. LITERATURE SURVEY

1. In this review, LINTA ANTONY and colleagues offer a comprehensive standalone structure for the anticipation of chronic kidney disease. The occurrence, prevalence, and progression of chronic kidney disease (CKD) have evolved over time across nations with diverse societal determinants of health. In most nations, the most frequent causes of chronic kidney disease (CKD) are diabetes and high blood pressure. Chronic kidney disease (CKD) is a condition marked by progressive declines in glomerular filtration rate (GFR) and signs of kidney damage. Younger deaths are more common among CKD patients. Since kidney damage can be prevented or even reversed by early discovery, doctors must promptly diagnose a variety of CKD- related illnesses. Treatment and patient care can be enhanced by early detection. There are not enough general practitioners or nephrologists in many distant clinics and hospitals to identify the symptoms.
2. In this work, ERIK ALONSO et al. propose an AI system for out-of-clinic heart failure heartbeat location. During out-of- hospital cardiac arrest (OHCA), automated pulse detection would facilitate the prompt diagnosis of cardiac arrest and the identification of spontaneous circulatory recovery. In this work, the majority of defibrillator signals, such as ECG and TI, were combined to produce a dependable pulse recognition system. The study's dataset comprised 1140 ECG and TI segments from 187 OHCA patients. An OHCA expert panel classified 348 of the segments as pulse less electrical activity (PEA) and 792 as pulse-generating rhythm (PR). Adaptive filtering was used to first extract the impedance circulation component and its first derivative from the TI. The different subband components and the denoised ECG were then produced using wavelet decomposition of the ECG.
3. JIONGMING QIN et al. identify chronic renal disease using machine learning. imply that Chronic Kidney Disease (CKD) is a worldwide health concern that causes a great number of deaths and numerous other ailments. Patients usually miss the early stages of chronic kidney disease (CKD) since there are no overt symptoms. Patients can receive timely treatment to reduce the progression of CKD if it is identified early. Doctors can successfully accomplish this goal with the help of machine learning models because of their rapid and accurate identifying capabilities. This research suggests a machine learning method for diagnosing chronic kidney disease. The CKD informational collection, which was acquired from the College of California, Irvine (UCI) AI repository, is lacking a great deal of information. By choosing many full samples with the most similar measurements, KNN imputation was used to process the missing data for each incomplete sample. In real-world medical situations, missing numbers are frequent since there are several reasons why people could overlook particular measurements.
4. A New Method for Medical Diagnosis Using Uneven Data Ensemble Learning LIU NA et al. Physicians and pathologists can prevent mistakes brought on by fatigue, stress, and other conditions by using machine learning (ML) approaches. Additionally, medical data may be analyzed more extensively and rapidly. However, an unbalanced dataset severely impairs the efficacy of typical machine learningBtechniques, such as classification, especially when it comes to identifying minority groups, even though these techniques were able to achieve acceptable classification accuracy when applied to medical diagnoses. In order to overcome the limitations of current classification techniques, this study proposes a unique ensemble learning paradigm with three stages for medical diagnosis with unbalanced data: data pre-processing, the training base classifier, and the final ensemble. In the initial stage of data pre- processing, I combine an extension of the Synthetic Minority Oversampling Technique (SMOTE) with the Cross-Validated Committees Filter (CVCF). To do well in the classification process, SMOTE can filter the noisy examples, balance the input instances, and synthesize the minority sample.
5. An Overview of Machine Learning Methods for Gene Expression Analysis in Cancer Prediction Strong machine learning algorithms, which are commonly employed to develop cancer prediction models based on linked gene expression and mutation data, were proposed by Mahmud Khalsanet and colleagues in this paper. This article claims that gene expression information from the liver, gallbladder, kidney, ovarian, breast, lung, and central nervous system has been utilized to stratify patients, detect tumors, and forecast survival. Furthermore, I offer an overview of the numerous cancer biomarker research. The survey examines a large body of machine learning-based cancer research, including methods to identify biomarker genes, classify and predict cancer, and analyze RNA-Seq, microarray, and other data.

III. RELATED WORK

Over time, there have been changes in the incidence, prevalence, and progression of chronic kidney disease (CKD), especially in countries with a variety of social determinants of health. The two main causes of CKD in the majority of nations are diabetes and high blood pressure. Global guidelines define chronic kidney disease (CKD) as a condition that causes kidney function to gradually decline, as shown by indicators

of kidney damage and glomerular filtration rate (GFR). Individuals who have chronic kidney disease are at risk of dying young. Since early detection can prevent or even reverse renal damage, doctors must recognize a variety of CKD-related disorders as soon as possible. Better treatment and patient care may result from early detection. Few nephrologists or general practitioners are available to diagnose the symptoms in many regional hospitals and clinics. Patients now have to wait longer for a diagnosis as a result of this. Therefore, this study makes the case that developing an intelligent system to divide patients into "CKD" or "Non-CKD" groups will help physicians handle several patients and make diagnoses more rapidly.

IV. METHODOLOGY

The CKD dataset, with all of its attributes, is the input. Unknown attributes and duplicate data are eliminated during pre- processing. Every one of the selected traits and attributes has been chosen. The Random Forest (RF) technique is used to enhance classification performance. Grades will be assigned to the f-measure, accuracy, recall, and precision. A visual depiction of these characteristics will be shown. They created a neural organization-based classifier and employed image enrolment to detect changes in the kidney's morphology using a large amount of CKD data and the model's accuracy on their test data. Furthermore, most earlier research used the CKD informative index from the UCI AI repository. This study investigates the diagnosis of chronic renal disease using machine learning (ML). Numerous classification jobs are benefiting from the effective use of machine learning algorithms, which have been essential in identifying abnormalities in a range of physiological data.

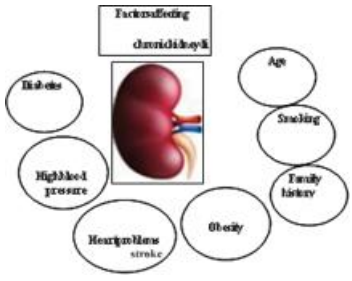

Figure .2.block diagram

A. **Loading The Dataset**

Every data analysis or machine learning project begins with loading data. The Chronic Kidney Disease (CKD) dataset is entered into the program in this instance. Each patient's CKD diagnosis outcomes and a range of factors are usually included in the dataset. You may access and manage the dataset by loading it, which enables you to perform further analysis.

B. Pre-Processing

By encoding data in a machine-readable manner, data pre-processing makes it possible to see a dataset as a group of objects. There can be missing values in the dataset. Missing values can be computed or eliminated from the dataset, but they must be handled with first. Usually, the mean, median, mode, or constant value of the feature is used to fill in the missing data. The categorical data of object type must be converted to float64 type since the object values are not usable for the analysis. The most frequent value in that column is used to replace null values in category attributes. By representing each distinct value as an integer, label encoding transforms categorical attributes into numerical values. As a result, the characteristics are automatically converted to integer type. The pandas software helps with pre-processing data.

C. Feature Selection

Finding and choosing the most pertinent attributes (features) from a dataset for model training is known as feature selection. Reducing dimensionality while preserving the most valuable features is the aim. The most crucial elements for CKD prediction can be identified using a variety of methods, including feature importance scores, mutual information, and correlation analysis. Selecting the appropriate characteristics can lower computational overhead and enhance model performance.

Table 1. Statistical analysis of the dataset of numerical features

Features	Mean	Standard	Max	Min
Age	51.483	17.21	90	2
Blood glucose random	148.037	76.583	490	22
Serum creatinine	3.072	4.512	76	0.4
Blood pressure	76.469	13.756	180	50
Bloodurea	57.426	49.987	391	1.5
Potassium	4.627	2.92	47	2.5
Packedcell volume	38.884	8.762	54	9
Sodium	137.529	9.908	163	4.5
Hemoglobin	12.526	2.815	17.8	3.1
Whiteblood cellcount	8406.12	2823.35	26400	2200
Red blood cellcount	4.707	0.89	8	2.1

D. Model Fitting and Testing

Data is split in half in a 4:1 ratio after feature selection, with 80% going toward training and 20% going toward testing learnt models. Random Forest is one of the suggested models that uses the data. The expected accuracy of each model is used to compare them.

E. Classification Performance

The ability of machine learning models to categorize or forecast outcomes, including the existence or absence of chronic kidney disease, is measured by classification performance. A number of metrics, including as accuracy, precision, recall, F1-score, and the confusion matrix, are frequently used to evaluate the performance of classification models. By dividing the number of accurately predicted occurrences by the total number of instances in the dataset, the accuracy ratio calculates how accurate the model is. In order to assess the model's capacity to prevent false positives, precision calculates the percentage of true positive predictions among all positive predictions. The model's ability to detect all pertinent situations is demonstrated by recall (sensitivity), which examines the percentage of true positive predictions among all real positive instances. A fair assessment of a model's performance is made possible by the F1- score, which is the harmonic mean of precision and recall.

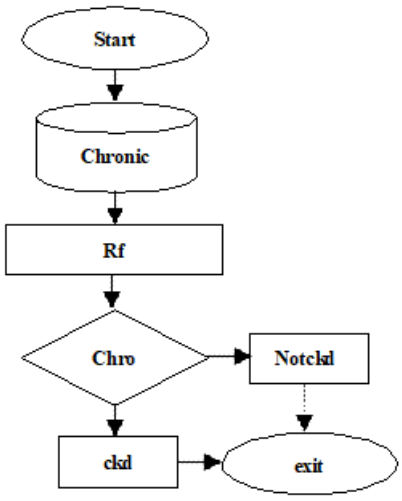

Figure 3. Flow diagram

V. ALGORITHM DETAILS

A. Random Forest Classifier

Like Decision Trees and Bagging Classifiers, Random Forest makes use of hyperparameters. Using the classifier- class of the random forest eliminates the need to combine a decision tree and a bagging classifier. The algorithm's repressor can be used for random forest regression challenges. The flexibility of the model is increased when a random forest is used throughout the tree creation process. When a hub is divided, it chooses the best part from a wide range of materials instead of the main component. There are several options available, most of which result in a better model. Only an irregular fraction of items in random timberland are examined in the calculation for isolating a hub. Consider using irregular boundaries for each component rather than looking for the best edges to make trees more unpredictable. Numerous decision trees are combined into a classification and regression ensemble using the Random Forest technique. Multiple decision trees are constructed using a random subset of the training data sets. The accuracy of the outcomes is increased by using multiple decision trees. This method can deal with missing data and has a short runtime. Instead of randomizing the training dataset, the process is randomized using Random Forest. Decision classes are produced via decision trees.

1. Here, the author first displays the pseudocode for Random Forest creation: Choose "K" features at random from the total "m" features, where k << m.
2. Determine the node "d" among the "K" characteristics by utilizing the optimal split point.
3. Use the optimal split to divide the node into daughter nodes.
4. Continue from step a to step c until the "l" number of nodes is attained.
5. To construct a forest, repeat steps a through d "n" times to produce "n" trees.

We'll make the prediction in the following step after the random forest classifier has been developed. The following is the pseudo code for random forest prediction:

1. Takes the test features and predicts the result using the rules of each decision tree that is randomly generated. The expected result (target) is then stored.
2. Determine the number of votes for every anticipated goal.
3. The random forest algorithm's highly voted anticipated target serves as the basis for the final forecast. Although the procedure is simple to comprehend, it is in some ways effective.

VI. RESULT ANALYSIS

According to the abstract, the table contrasts three classifiers—K-Nearest Neighbors (KNN), Random Forest

(RF), and Random Forest—based on how well they predict outcomes in the context of chronic renal illness. The percentage of successfully identified cases is measured by the accuracy statistic, and RF's 98.33% and 96.67% accuracies, respectively, show how well they can identify patterns in the dataset. With scores of 100% and 92%, respectively, and percentages of 97.37% and 94.74%, precision—a gauge of the classifier's capacity to avoid false positives—shows how resilient RF is. With scores of 98.67% and 97.3%, respectively, the F1-score—a harmonic mean of precision and recall—validates RF's overall superior performance. These results highlight RF's capacity to offer trustworthy and precise prognoses for chronic renal disease, which makes them attractive options for additional research and possible clinical use.

Table .2.comparison table

Classifiers	KNN	RF
Accuracy%	96.67	98.33
Precision%	92	100
Recall%	94.74	97.37
F1-score%	97.3	98.67

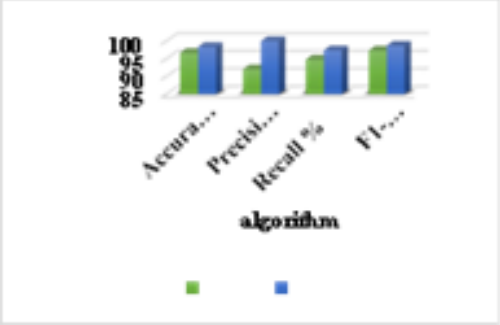

Figure.4. comparison graph

VII. CONCLUSION

Last but not least, this study aims to predict chronic kidney disease, which is on the rise, using machine learning. This work aims to advance healthcare decision- making tools by evaluating and comparing Random Forest (RF) approaches. In order to overcome the difficulties presented by the progressive nature of chronic renal illness, early detection and monitoring are essential. A major step toward reducing the impact of this health burden on people and healthcare systems has been taken with the findings and insights from this work, which show promise in enhancing predictive capacities and influencing successful therapies.

VIII. FUTURE WORK

In order to forecast chronic renal disease, future research in this field should focus on enhancing and growing the machine learning models outlined in this work. Predicted accuracy and robustness could be increased with more research into sophisticated algorithms and ensemble techniques. Additionally, merging diverse and sizable datasets would enhance model generalization and practical implementation. The models' ability to represent the intricacies of chronic kidney disease risk may be improved by include more pertinent variables, such as lifestyle factors or genetic markers.

REFERENCES

[1] J. M. Bolarn, f. Cavas, j. S. Velázquez, j. L. Alió Sci., "a total solo system for persistent renal sickness expectation," Appl., vol. 10, no. 5, pp. 1874, March 2020.

[2] "A machine learning framework for the detection of pulses during cardiac arrest outside of a hospital," by M. Abdar, M. Zomorodi-moghadam, x. Zhou, r. Gururajan, x. Tao, and p. D. Barua, et al., pattern recognition letters, vol. 132, pp. 123-131, April 2020.

[3] "A method for identifying chronic renal disease using machine learning," J. Med., 42, no. 12, p. 261, 2020; S. Thompson, J.

[4] Mather, C. M. Waszynski, and R. S. Dicks, with J. P. Corradi.

[5] A new technique for using unbalanced data in ensemble learning for medical diagnosis, R. A. Shah, M. Zaman, S. A. Khan, N. A. Ganai, M. Ashraf, and others, Springer Singapore, pages. Pages 239–255, 2021.

[6] Resuscitation, Vol. 5, 147, pages, February 2020; H. Kwok, J. Coult, J. Blackwood, S. Bhandari, P. Kudenchuk, and T. Rea.

[7] W. Zhu, L. Xie, J. Han, and X. Guo, "Analysis for Disease Gene Association Using Machine Learning," Cancers, vol. 12, no. 3, pp. 603, March 2022.

[8] The International Journal of Computer Applications, 43(6), 524-536, July 2020, "Covid-19 future forecasting using supervised machine learning models" by Jain, D. and Singh, V.

[9] Using corneal imaging data and machine learning to detect keratoconus. Geneva, Switzerland, 2021.

[10] F. Petropoulos and S. Makridakis, "Diabetes prediction using assembly of different machine learning classifiers," PLoS ONE, vol. 15, no. 3, March 2022.

[11] The international journal of computational neuroscience and engineering published an article titled "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection

techniques" in June 2021. S. Kumari, D. Kumar, and M. Mittal wrote it.